Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application

Thomas M. Hamill and Jeffrey S. Whitaker

NOAA Earth Systems Research Lab, Physical Sciences Division
(formerly the Climate Diagnostics Center)
Boulder, Colorado

Submitted to *Monthly Weather Review*

14 October 2005

Corresponding author address

Dr. Thomas M. Hamill
NOAA ESRL/PSD,
R/PSD 1,
325 Broadway
Boulder, Colorado 80305-3328
E-mail: tom.hamill@noaa.gov
Phone: 1 (303) 497-3060
Fax: 1 (303) 497-6449

ABSTRACT

A general theory is proposed for the statistical correction of weather forecasts based on observed analogs. An estimate is sought for the probability density function (pdf) of the observed state, given today's numerical forecast. Assume that an infinite set of reforecasts (hindcasts) and associated observations are available, and that the climate is stable. Assume that it is possible to find a set of past model forecast states that are nearly identical to the current forecast state. With the dates of these past forecasts, the asymptotically correct probabilistic forecast can be formed from the distribution of observed states on those dates.

Unfortunately, this general theory of analogs is not useful for estimating the global pdf with a limited set of reforecasts, for the chance of finding even one effectively identical forecast analog in that limited set is vanishingly small. Nonetheless, approximations can be made to this theory to make it useful for statistically correcting weather forecasts. For instance, when estimating the state in a local region, choose the forecast analogs only based on the local weather, for which there usually is an ample supply with a modest-sized reforecast.

Several approximate analog techniques are then tested for their ability to skillfully calibrate 24-h accumulated probabilistic quantitative precipitation forecasts (PQPFs). A 25-year set of reforecasts from a reduced-resolution global forecast model is used. The analog techniques find past ensemble-mean forecasts in a local region that are similar to today's ensemble-mean forecasts in that region. Probabilistic forecasts are formed from the observed weather on the dates of the past analogs. All of the analogs techniques provide dramatic improvements in the Brier skill score relative to basing probabilities on

the raw ensemble counts or the counts corrected for bias. However, the analog techniques were not much more skillful than those from a logistic regression technique. Among the analog techniques tested, it was determined that small improvements to the baseline analog technique that matches ensemble-mean precipitation forecasts are possible. Forecast skill can be improved slightly by matching the ranks of the mean forecasts rather than the raw mean forecasts, by using highly localized search regions for shorter-term forecasts and larger search regions for longer forecasts, by matching precipitable water in addition to precipitation amount, and by spatially smoothing the probabilities.

1. **Introduction**

Despite much recent progress in numerical weather prediction, weather forecasts are still subject to error, both as a result of the growth of initial-condition errors and model errors. Near-surface forecasts and forecasts of hydrologic variables such as precipitation or cloud properties are particularly error prone, in part because these physical processes often occur at scales below those resolved by the model. These effects must be "parameterized," and developing accurate parameterizations is a difficult endeavor. As computational power increases, forecast models have been updated and increased in resolution to address these problems.

A complementary pathway to improved forecasts for users is to utilize a known weather forecast model consistently, so that a long past time series of weather forecasts are available. If the climate is relatively stable, then the errors in past similar weather scenarios can be used to statistically correct the current numerical forecast. This approach is of course well established, being the essence of Model Output Statistics, or "MOS" techniques (Glahn and Lowry 1972, Carter et al. 1989). If today's numerical forecast indicates relatively ordinary conditions, then perhaps the past few months or year will have exhibited enough other similar scenarios that the current forecast can be properly adjusted. But what if the weather is relatively unusual? Suppose high rain amounts are forecast for a desert location; it is likely that there will have been few similar forecast events at that location that can be used to determine how to correct the forecast. If a model's systematic errors are similar throughout a region, then the effective sample size can be increased by pooling the training data over many geographic locations. However,

if the forecast errors are regionally dependent, there may be no effective substitute for a training database that spans many years or decades.

The presumed benefit of large training datasets motivated our foray into "reforecasting," the production of a large data set of retrospective forecasts using the same model that is run operationally. Recently, we produced a sample reforecast data set (Hamill et al. 2004 and Hamill et al. 2005, hereafter HWM05). The novel feature of this prototype data set was the extraordinary length and volume of the reforecast training data set, 25+ years of 2-week ensemble forecasts initially centered on a reanalysis state. Such a large training data set may permit accurate statistical adjustments even for some relatively rare events. A disadvantage of relying on reforecasts is the computational expense of generating them. To reduce this expense, we used a 1998 version of the National Center for Environmental Prediction's (NCEP's) Global Forecast System (GFS) at a reduced, T62 resolution; certainly, it would be preferable to use a newer, higher-resolution model.

In the HWM05 article, a simple, skillful, two-step analog statistical correction technique was introduced as a way of making probabilistic forecasts, and we return to consider this analog technique more closely here. The first step in the analog technique was to compare the current forecast to all past forecasts at a similar time of the year in a local region. Second, the dates of the closest matches were determined, and an ensemble was formed from the higher-resolution observed weather on those dates, from which probabilities could be calculated from the event frequency. A gridded field of probabilities was produced, tiling together the probabilities computed for each independent region. Probabilistic forecasts from this ensemble were both reliable and

specific when compared to forecasts generated from a more recent, higher-resolution version of the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) ensemble. The analog technique was able to correct the forecast bias and ensemble spread deficiencies and downscale the output to the scale of the higher-resolution, 32-km precipitation analysis (the North American Regional Reanalysis, Mesinger et al. 2005).

As HWM05 was an overview article, much detail and context were missing, and this article is intended to provide this context. Specifically, the purpose of this article is: (1) to provide an underlying theoretical basis for use of analog technique and explain the practical approximations that must be made in order to apply it, (2) to compare the analog technique against a few logical alternatives, such as logistic regression, and (3) to explore whether the simple analog technique of HWM05 can be enhanced further through slight algorithmic variations. The intent of this article, however, is not to provide an exhaustive comparison of the myriad of possible calibration techniques that exist in the literature. The few non-analog methods we test are included primarily to help understand the reasons why the analog methods provide such an improvement over probabilities set from the raw ensemble forecasts. A rigorous comparison against other calibration techniques would indeed be interesting, but many of them were designed with small training data sets in mind, applying approximations such as compositing training data over many locations.

Below, we first provide a theoretical underpinning for this two-step analog technique (section 2). The data sets and a variety of specific statistical correction

techniques are then described (section 3), and an intercomparison of these techniques are provided (section 4), with conclusions (section 5).

The data used in this study is also freely available (section 3.a), and we encourage others that would be interested in testing their methods to use this data set and compare their results against the benchmarks set here.

2. **Theoretical basis of the analog technique and simplifying assumptions.**

Let us suppose we have an ensemble of gridded forecast model states for a particular time. Assume that there are *n* components to the state vector, and *m* ensemble members. We thus have a *m***n*-component forecast vector $\mathbf{x}^f$ composited from the ensemble members' forecasts:

$$\mathbf{x}^f = \left(x_1^f(1), \ldots, x_1^f(m), \ldots, x_n^f(1), \ldots, x_n^f(m) \right) = \left(\mathbf{x}_1^f, \ldots, \mathbf{x}_n^f \right) \quad (1)$$

Suppose we are interested in the *p*-dimensional observed state of the atmosphere

$$\mathbf{x}^t = \left(x_1^t, \ldots, x_p^t \right) \quad (2)$$

at the same time; this could be the state at grid points or specific locations. The probabilistic weather forecast problem is then conceptually simple; we seek

$$f\left(\mathbf{x}^t \right) \Big| \mathbf{x}^f \quad (3)$$

where $f(\cdot)$ denotes the probability density function; that is, we want to accurately quantify the probability distribution of the observed state of the atmosphere, given the ensemble forecast. Were the observed state comprised of the same variables at the same locations as the forecast state and the forecast model were perfect (i.e., chaos was the only source of error, and the ensemble perfectly represented this), then the relative frequency from the ensemble would provide an adequate definition of any event probability, accurate within sampling error:

$$P\left(x_i^t > T\right) = \frac{1}{m}\sum_{j=1}^{m} I\left(x_i^f(j), T\right) \tag{4}$$

where T is the threshold for some chosen event, $I\left(x_i^f(j), T\right) = 1$ when $x_i^f(j) > T$, and 0 otherwise. Unfortunately, ensemble forecasts are typically quite imperfect, due to model errors and deficiencies in the method of constructing the ensemble.

If the climate was stable and it was possible to compute a nearly infinite set of reforecasts with associated verification data, then it would be possible to compute eq. (3) directly, even in the presence of model error. With this nearly infinite ensemble, we could simply find past forecast states that were almost identical to the current forecast state and then determine eq. (3) from the distribution of the observed states on those dates. Suppose there are s reforecasts of the same forecast lead time that are practically identical to the current forecast at that lead. Let $\mathbf{x}^{t|r} = \left(\mathbf{x}^{t|r}(1), \ldots, \mathbf{x}^{t|r}(s)\right)$ denote the collection of the s associated past observed states on the dates of the nearly identical reforecast analogs. Then to find the event probability at a given location,

$$P\left(x_i^t > T\right) = \frac{1}{s}\sum_{k=1}^{s} I\left(x_i^{t|r}, T\right) \quad , \tag{5}$$

where $I\left(x_i^{t|r}, T\right)$=1 when $x_i^{t|r}(j) > T$ and 0 otherwise. All that is being done here is to determine the fraction of time when the threshold is exceeded using the observed data associated with the chosen analogs. If the observed state is actually providing describing the atmospheric state at much smaller scales than the original forecast, then this procedure amounts to a statistical downscaling (Zorita and von Storch 1999).

This process of eq. (5) is conceptually illustrated in Fig. 1 with a synthetic 10000-day reforecast data set. Here we have created a time series of reforecast and associated observed data; the state is a scalar, and the forecast is deterministic, so the problem can be visualized two dimensionally. Consider the event that the true state is > 0.0. Suppose our criteria for closeness of a reforecast to the current day's forecast was to be within a window of 0.5 units. To apply eq. (5), we find the forecast points in vertical columns of this width and then count the fraction with observed data > 0.0; the horizontal bars in Fig. 1 provide the probability based on this simple count. Also plotted is a fitted logistic regression curve (Wilks 1995) to the data. In comparison to the analog process, the logistic regression parameters are fit using all of the scatter plot data at once, rather than just the data from close forecast analogs. Depending on the data, the smooth, S-shaped logistic-regression curve may provide a better or worse fit than the analogs when sample size is finite.

It is worth considering the asymptotic error characteristics of such a forecast approach as skill increases or decreases. If the forecast is totally uncorrelated with the observed data, then using (5) will reproduce the climatological distribution, within

sampling error. If the forecast system's fidelity improves so that the correlation of forecast and observed approaches 1.0, the probabilistic forecasts will get increasingly sharp without losing reliability. In the asymptotic limit that the forecast error approaches zero, the probabilistic forecast will approach a perfect deterministic forecast. In this case, of course, a reforecast would be unnecessary, but as this asymptotic limit is only of theoretical concern (Lorenz 1963), it is at least comforting to know that the performance of the statistical analog approach in eq. (5) will improve as the forecast model improves.

The analog process is quite simple as illustrated in Fig. 1, but suppose the model state is a 100-member ensemble forecast of winds, temperatures, humidity, and geopotential at millions of grid points covering the globe. Even with billions of years of reforecasts, it may prove difficult to find many close global analogs (Lorenz 1993, p. 86, Van den Dool 1994). And even were a reforecast available over such a long period of time, the climate, and indeed the continents themselves, were not very stable. Hence, simplifying assumptions are required. Some possible assumptions may include:

- If we are concerned specifically with the assessing the probabilities at a particular location, only the forecast model state around that location may be needed, e.g., to estimate probabilities for Washington, D.C., it is only necessary to find the dates of past forecasts matching today's D.C.-area forecast; matching or not matching at other distant locations is irrelevant (Van den Dool 1989). In the terminology of linear regression, the model forecast state at the distant locations would not make useful predictors.

• If provided with a forecast ensemble, it may be unnecessary to match all the aspects of the ensemble; matching the mean state may be sufficient, or perhaps the mean and the spread, rather than requiring that each member match.

• If considering an event like surface temperature, it may be sufficient to match reforecasts of surface temperature alone, ignoring other forecast aspects such as upper-level winds or temperatures.

3. **Data sets and methods.**

Below, we provide a brief description of the reforecast and verification data sets, and then more detail on various statistical correction techniques and the metrics for evaluating forecast skill. Our focus is on the calibration of 24-h accumulated probabilistic quantitative precipitation forecasts at the scale of the scale of the verification data, ~ 32 km.

a. Reforecast and verification data sets.

HWM05 provide a more complete description of the reforecast data set. The forecast model is a 28-level, T62 resolution version of NCEP's Global Forecasting System (GFS) model using physics that were operational in the 1998 version of the model. The reforecasts were generated at the NOAA lab in Boulder, Colorado, and real-time forecasts are now generated at NCEP and archived in Boulder. A 15-member ensemble was produced every day from 1979 to current, starting from 0000 UTC initial conditions. The ensemble initial conditions consisted of a control initialized with the NCEP-National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al.

1996) and a set of 7 bred pairs of initial conditions (Toth and Kalnay 1993, 1997) re-centered each day on the reanalysis initial condition. The breeding method was the same as that used operationally in January, 1998. The forecasts extended to 15 days lead, with data archived every 12 h. Winds, temperature, and geopotential height are available at the 850, 700, 500, 250, and 150 hPa levels. 10-m wind components, 2-m temperature, mean sea-level pressure, accumulated precipitation, convective heating, precipitable water, and 700 hPa relative humidity were also archived. General reforecast data is available online at www.cdc.noaa.gov/reforecast .

The observed 24-h precipitation data was taken from the North American Regional Reanalysis, or NARR, described in Mesinger et al. (2005). The data was on a 32-km Lambert-conformal grid. Only grid points over the conterminous U.S. were used.

For all subsequent experiments, the data set will consist of reforecasts from 1 January 1979 to 31 December 2003, 25 years of forecasts.

For those who wish to test their own methods against those described here, the specific reforecast and training data used in this experiment is available at www.cdc.noaa.gov/reforecast/testdata.html .

b. Probabilistic estimation techniques.

We now briefly review 10 different techniques for estimating precipitation event probabilities. The first does not use the reforecast data set; the rest do.

1) ENSEMBLE RELATIVE FREQUENCY

The simplest approach uses no statistical calibration. The relative frequency of event occurrence is estimated directly from the 15-member ensemble, interpolated to the

32-km NARR grid. For example, if 5 of the 15 members at a point indicate greater than 25 mm rainfall, the probability is set to 33.3 percent.

2) BIAS-CORRECTED RELATIVE FREQUENCY

In this procedure, probabilistic forecasts are generated from an ensemble of forecasts, where each member has been bias-corrected according to the long-term bias statistics for that grid point and time of year. This follows a technique proposed by Y. Zhu at NCEP (personal communication, 2005). Let $F_Y^C(y)$ denote the cumulative distribution function CDF of the 24-h precipitation amount of climatology, defined by

$$F_Y^C(y) = Pr_Y^C(Y \le y). \tag{6}$$

Here Y is the random variable, y the specific amount being considered, and $Pr(\cdot)$ indicates the probability, which will be determined by frequency from a large sample. Similarly, define a CDF for the 24-h ensemble forecast amount, $F_X^E(x)$ defined by

$$F_X^E(x) = Pr_X^E(X \le x). \tag{7}$$

The technique is then rather simple. For a given day of the year, we compute $F_Y^C(y)$ and $F_X^E(x)$ using the 25 years x 91 days (centered on the day of interest) of analyzed precipitation and interpolated member forecasts. Then, for a given ensemble forecast on that day with the value x, we determine a value y such that $F_Y^C(y) = F_X^E(x)$. The ensemble member forecast x is then replaced with the value y. This is illustrated in Fig. 2 for today's hypothetical forecast and the forecast and observed CDFs. As implemented at

NCEP in 2004, the technique is slightly different; only a short training sample is used, such as the past 30 days, and the technique generates CDFs that are not location-specific, instead representing an average over the domain.

3) Basic Analog Technique

This procedure was described in HWM05, and a simple pictorial representation of the method is provided in Fig. 3. As suggested in section 2, application of the full analog theory assumes a nearly infinite training sample. Without this, we adopted several of the simplifying assumptions; namely, we search only for local analogs, match the ensemble-mean fields, and consider only the model forecast of precipitation (no winds, temperature, geopotential, etc.) in selecting analogs.

The first step of the procedure is to find the closest local reforecast analogs to the current numerical forecast. That is, within a limited-size region, the forecast for the day under consideration (the map in the top row in Fig. 3) is compared against past forecasts in that same region and at the same forecast lead. Specifically, the ensemble-mean precipitation forecast pattern is computed at a subset of 16 coarse-mesh reforecast grid points surrounding the region where analogs are sought. The ensemble-mean forecasts at these points are compared to ensemble-mean reforecasts at these points in all the other years (a cross-validation procedure), but only those within a window of 91 days (+ / - 45 day window) around the date of the forecast. This is done under the presumption that model biases may change substantially with the time of year. The root-mean square (RMS) difference between the current forecast and each reforecast is then computed, averaged over the 16 grid points. The *n* historical dates with the smallest RMS difference

are chosen as the dates of the analogs (the maps in the second row of Fig. 3). The next step is the collection of the ensemble of observed weather on the dates of the closest *n* analogs (third row in Fig. 3). Probabilistic forecasts are then generated using the ensemble relative frequency; for example, if 3/4 of the members at a grid point had greater than 10 mm accumulated rain, the probability of exceeding 10 mm was set to 75% (fourth row in Fig. 3). This can then be compared to the observed weather (bottom row). The process is then repeated for other locations around the US. A 32-km probabilistic forecast is generated, tiling together the local analog forecasts.

Commonly, many more than the four members shown in this figure would be used. In actuality, ensembles of size 10, 25, 50, and 75 were computed. In subsequent figures, the skill scores will be plotted only for the optimal size. In general, the optimal size was smaller for heavier precipitation events and shorter forecast leads. For more information on the optimal ensemble size, see HWM05, Fig. 7.

4) LOGISTIC REGRESSION

Logistic regression estimates event probabilities at a particular location through an equation of the form

$$P(x_i^t > T) = \frac{1}{1 + exp\left(\widehat{\beta}_0 + \widehat{\beta}_1\phi_1 + \ldots + \widehat{\beta}_n\phi_n\right)} \tag{8}$$

where $\hat{\beta}_0, \ldots, \hat{\beta}_n$ are fitted regression coefficients and $\phi_1, \ldots, \phi_n$ are model predictors based on the forecast data. Here *n*=2, ϕ_1 is the square-root of the ensemble-mean precipitation amount interpolated to the observation location (the square-root transformation made the data have less of a heavy right tail), and ϕ_2 is the column

ensemble-mean precipitable water interpolated to the observation location, measured in mm. The regression coefficients are determined in a cross-validated manner; when coefficients are developed for grid points for a particular year, that year's forecasts are excluded from the training data. As with the basic analog technique, a 91-day window of forecasts is used. Multiplied by the 24 years of training data, this provided a sample size of 2184. Logistic regression techniques were tried without precipitable water, and without the power transformation; both were somewhat less skillful, and results for these will not be presented.

Unlike the analog technique, the logistic regression technique uses all of the data to determine the regression curve, not just a subset of close analogs. This has the advantage of increasing the sample size but the potential disadvantage that cases very dissimilar to the forecast of interest on that day are also being used to estimate the probabilities. Another disadvantage of the logistic regression technique is it provides only a probability for the event threshold under consideration; if probabilities are desired for a different threshold, the regression analysis must be repeated. In comparison, when using analog techniques, once the analog members are chosen, probabilities can be defined quickly for any event threshold.

5) Basic Technique Using Individual Members

This technique is similar to the technique in section 3.b.3, but instead of searching for analogs of the ensemble mean, member 1's forecast is compared against past member 1 reforecasts, the dates of the 5 closest matches were noted, and the process was repeated for the rest of the 15 members, producing 75 (possibly non-unique) dates. If an ensemble

of size smaller than 75 is to be formed, the dates of the closest pattern matches from the set of 75 dates are used. Comparing the skill of these forecasts to that of the basic technique will indicate whether the information content can be distilled down to the ensemble mean, or whether there was extra information in each member.

6) BASIC TECHNIQUE INCLUDING PRECIPITABLE WATER

This technique repeats the basic analog technique from section 3.b.3, but instead of only matching the ensemble-mean precipitation amount, column precipitable water is included as well. When measuring the closeness of a past reforecast, the ensemble-mean precipitation is weighted by 70 percent and the precipitable water by 30 percent.

7) BASIC TECHNIQUE INCLUDING 2-M TEMPERATURE and 10-M WINDS

This technique repeats the basic analog technique from section 3.b.3, but instead of only matching the ensemble-mean precipitation amount (in mm), 2-m temperature (K) and 10-m *u*- and *v*-components of the wind (ms^{-1}) are included as well. When measuring the closeness of a past reforecast, the differences in precipitation, temperature, and wind speed between today's forecast and the past reforecasts at the 16 grid points are squared and then summed, and the analogs' dates are those with the smallest sums.

8) RANK ANALOG TECHNIQUE

This technique is generally the same as the basic analog technique described in section 3.b.3, with one exception. When determining the closest matches, at each of the 16 grid points, the rank is computed for today's precipitation forecast amount when

pooled with the reforecasts. Similarly, the rank of the precipitation amount is determined at each grid point for each of the candidate reforecasts. The analog dates are those with the lowest sum of the absolute value of rank differences over the sixteen points.

The rank analog technique is included here because results will show (section 4, Fig. 13) that at early leads, the basic analog technique produced somewhat unreliable forecasts, under-forecasting precipitation probabilities. Upon closer examination, it was determined that this was primarily because the distribution of precipitation forecast amounts was skewed, with lighter amounts more common than heavier amounts. Consequently, the basic technique's closest forecast analogs more commonly had slightly less precipitation than today's forecast more often than they had slightly more precipitation. Using a rank-based approach was proposed as a way of ensuring that more equal numbers of heavier and lighter forecast events were used as analogs.

9) RANK ANALOG WITH SMALLER SEARCH REGION

This technique repeats the rank analog technique from section 3.b.8, but instead of finding a match over a set of 16 grid points (see Fig. 3), only the four center grid points are used in determining the dates of the reforecast analogs. A comparison of this against the rank analog will indicate whether the size of the search region is an important determinant of forecast skill.

10) SMOOTHED RANK ANALOG TECHNIQUE.

Most of the prior analog approaches discussed in this article produce probability estimates for an 8x8 box of 32-km grid points, finding analogs using the surrounding,

large-scale forecast fields (see Fig. 3). To produce a national map of the probabilities, the process is repeated for other regions, and the final map is a composite of the 8x8 patches. Unfortunately, sometimes the dates of the analogs can be quite different for one set of 8x8 boxes when compared to its adjacent sets. This may result in a slight discontinuity of the probabilities at the boundaries between patches. Accordingly, we test a simple smoothing algorithm to eliminate this effect. Aside from the smoothing applied at the end, this method will be identical to the rank analog technique, described previously.

To understand the smoothing, consider Fig. 4. Say we seek to estimate the probabilities on the 32-km grid at the orange grid points. The analogs were determined by matching today's forecast at the large-scale grid points (large black dots) to past forecasts at these same dots. For the orange dots, there were actually nine separate regions where analogs and the subsequent probabilities were calculated that overlapped the orange dots, shown in the nine panels of Fig. 4. In all previously described analog algorithms, estimates at eight of these were thrown away, and only the probabilities from analogs from the middle search region were used.

Here the smoothing algorithm uses several of the estimated. Consider the orange grid point surrounded by the red box in Fig. 4. Let w_{ul}, w_{um}, w_{ur}, w_{ml}, w_m, w_{mr}, w_{ll}, w_{lm}, and w_{lr}, denote the weights applied to the probability estimates using the upper left box, the upper middle box, and so on. Let d_{ul}, d_{um}, d_{ur}, d_{ml}, d_m, d_{mr}, d_{ll}, d_{lm}, and d_{lr} indicate the distance between the center point of each calculation region (the blue dot) and the red box, and let

w'_{ul}, w'_{um}, w'_{ur}, w'_{ml}, w'_m, w'_{mr}, w'_{ll}, w'_{lm}, and w'_{lr} represent a non-normalized weight. Here, define the threshold distance D to be $\sqrt{128}$, the distance in NARR grid points

between the upper-left and middle blue dots. First, a non-normalized weight is calculated according to

$$w'_{xx} = \begin{cases} \dfrac{D - d_{xx}}{D + d_{xx}} & if\ d_{xx} < D \\ 0 & if\ d_{xx} \geq D \end{cases} \tag{9}$$

where *xx* is one of the nine regions, e.g., *ul*. After all nine non-normalized weights are calculated, then weights are normalized by the sum of the non-normalized weights. For example,

$$w_{ul} = \frac{w'_{ul}}{w'_{ul} + w'_{um} + w'_{ur} + w'_{ml} + w'_{m} + w'_{mr} + w'_{ll} + w'_{lm} + w'_{lr}} \tag{10}$$

For example, the nine weights for the red box from upper left to the lower right are 0.134, 0.188, 0.000, 0.261, 0.379, 0.004, 0.004, 0.029, and 0.000.

c. Performance measures.

A primary metric of forecast performance will be the Brier Skill Score (BSS; Wilks 1995). A Brier Score (ibid) was calculated both for the forecast (BS_f) and for climatology (BS_c), and the BSS was then calculated according to

$$\text{BSS} = 1.0 - BS_f / BS_c \ , \tag{11}$$

Climatology here was determined from the sample event probability as computed from the observed 1979-2003 data. To ameliorate the possibility that false skill was reported from using a composite climatology over seasons or diverse locations (Hamill and Juras 2005), climatological probabilities were determined separately for each grid point and

day of the year using the 25 years times 61 days (the day of interest +/- 30 days). Climatological probabilities were not cross validated.

Because of the extremely large sample size of the forecasts, even small differences in the BSS tended to be statistically significant. Tests of significance were evaluated with the block bootstrap technique described in Hamill (1999).

Reliability diagrams (Wilks 1995) will also be used to illustrate the degree of correspondence between forecast probabilities and observed relative frequencies.

4. **Results**.

Tables 1 and 2 provide the BSS for each of the 10 techniques discussed in the previous section. The bias-corrected relative frequency technique improved the forecast of 2.5-mm forecasts somewhat compared to the ensemble relative frequency technique, but it tended to lower the skill of the 25-mm forecasts. All the rest of the methods provided a consistent, very large improvement over the ensemble relative frequency technique. With the exception of the bias-corrected relative frequency, the skill differences between the various reforecast-based calibration techniques were much smaller; having achieved most of the skill with the basic analog technique or the logistic regression technique, other methods had only slightly larger or smaller skill scores.

Let's examine these results in more detail. First, consider forecasts from the ensemble relative frequency technique. Figure 5 provides a plot of the BSS of this technique as a function of the time of the year and the forecast lead time. Unsurprisingly, skill was larger at short leads and larger in the cool season. Regrettably, many of the forecasts were less skillful than the reference seasonal climatological distribution. Light

precipitation forecasts in April were particularly unskillful; a subsequent examination of reliability diagrams, not shown, showed that light amounts were over-forecast in April much more commonly than at other times of the year. Note also that skill increased when verified on coarser-spaced grids (not shown), as indicated by Gallus (2002).

The bias-corrected relative frequency technique improved the cool-season precipitation forecasts but tended to make the warm-season forecasts even less skillful (Fig. 6). How could this bias-correction technique worsen the forecast? To understand this, we chose an 8x8 set of 32-km NARR grid points centered in northern Mississippi and examined the 1-day forecasts in mid-August, a date and location where the skill of 25-mm forecasts decreased. Figure 7a shows the average CDFs for the forecast and observed over these grid points. Light precipitation events were forecast much too frequently; for example, a 2-mm forecast was approximately at the 38th percentile of the forecast's cumulative distribution, while the 38th percentile of the observed distribution was 0 mm. Conversely, precipitation events above 16 mm were forecast less commonly then they were observed. Figure 7b provides a scatter plot of 1-day forecasts of a single member from the ensemble, plotted against the observed precipitation. Plotted over top, the green line illustrates the function that relates forecast precipitation to its adjusted amount based on the CDF differences. The orange line indicates the mean of the conditional distribution of the observations given the forecast, plotted using a running-line smoother with a window width of 4 mm (Hastie and Tibshirani, 1990). Using the CDF adjustment, all forecasts below ~5 mm were adjusted to zero precipitation, while a forecast precipitation amount of 30 mm was adjusted to ~ 45 mm. Figure 7c illustrates the pdf adjustment for an ensemble forecast with a mean of ~25 mm. The dashed

histogram indicates the frequency distribution of the adjusted ensemble forecast, while the red histogram indicates the conditional distribution of observations, given an ensemble-mean forecast of between 23 and 27 mm. As can be seen, the adjustment shifted the distribution further away from the conditional distribution of observations, so averaged over many similar cases, these forecasts should have scored worse than the uncorrected forecast.

At first glance, the difference between the CDF adjustment and the adjustment implied by the conditional distribution of observations appear contradictory: if there were truly more high-precipitation events in the observed CDF than in the forecast CDF, then why would the conditional distribution of observed events given the 25-mm forecast have a mean observed value lower than the mean forecast? The discrepancy was due to the lack of a strong relationship between forecast and observed data, i.e., the largest observed amount in the sample pool did not occur when the forecast the was largest. Had the forecasts and observations been very highly related, then the conditional distribution of observed events given a 25-mm ensemble-mean forecast would indeed be larger than 25 mm, far different than the unconditional climatology. In fact, as shown by the difference between the green and orange lines, the CDF-bias correction was in the wrong direction for forecasts above ~ 17 mm, where the CDF correction suggested a mapping of the forecast to higher amounts while the mean observed given the forecasts indicated the preferred mapping was toward lower amounts. Asymptotically, then, the performance of this bias correction based upon differences in the CDFs was likely to make already bad forecasts (i.e., deficient in spread and poorly correlated with observations) worse when the observed and forecast CDFs differed. The stronger the forecast-observed relation, the

more one can expect the CDF-based bias correction to improve forecast skill. However, even then the CDF-adjustment method by construction did not correct for spread deficiencies, so even with a perfect forecast-observation relationship, this calibration technique may not result in as skillful probabilistic forecasts as other methods designed to address that aspect as well.

Suppose a bias correction was based on some type of regression analysis rather than the CDF technique. Then, if forecasts and observed were uncorrelated, all member forecasts would be adjusted to the climatological mean, regardless of their initial value, converting the ensemble into a deterministic forecast. And were the mean of the forecast distribution shifted but the spread of the ensemble preserved, one could envision situations where the shift could create members with unmeteorological, negative-valued precipitation amounts. The overall lesson seems to be that it will be difficult to improve probabilistic forecast through some simple bias adjustments; errors in the mean and in the spread should both be addressed.

We return to examining the rest of the correction methods, all of which did improve upon the BSS compared to the ensemble relative frequency. Figure 8 shows the BSS of the basic analog approach for each month. The forecasts were almost universally skillful relative to climatology, with more skill in the cool season and more skill at short leads and lesser amount thresholds.

Figure 9 shows that the logistic regression technique performed quite similarly. The forecasts were slightly more skillful than the basic analog forecasts at day 1, and generally similar or less skillful than the basic analog at longer leads. In Fig. 9, this skill comparison is visualized through the use of the grey shading. When the shading was

above the plotted line for the logistic regression technique, this indicated that the basic analog technique had a proportionally higher BSS; when the shading was below the logistic regression line, the basic analog technique had a lower BSS. Because of the large sample size, even small differences tended to be statistically significant. Applying a block bootstrap technique (Hamill 1999), the magnitude of yearly differences that were statistically significant at the 95 percent confidence level are presented in the last row of Tables 1 and 2; monthly differences that were significant were typically 2 or 3 times larger.

Was there an advantage to fitting individual ensemble members rather than the ensemble mean? Figure 10 presents the BSS for the basic technique using individual members, along with a comparison of skill relative to the basic analog technique. Fitting individual members provided forecasts of approximately equal skill for short leads, but for longer-range forecasts in the cool season, the skill was considerably worse when fitting individual members. We hypothesize that at longer leads in the cool season, the filtering properties of the ensemble mean were helpful in extracting the predictable signal obscured by the chaotic error growth; when fitting individual members, one was fitting more noise than signal at the longer leads. In the summer, we hypothesize that model systematic errors played a more dominant role in limiting forecast skill, and that the role of chaotic error growth was secondary.

The logistic regression technique included an extra predictor for precipitable water. If this extra predictor were incorporated into the analog technique, would the skill improve as well? Figure 11 presents the skill of the basic technique including precipitable water. While warm-season forecasts of light precipitation amounts were

improved somewhat, otherwise the skill of the two methods were very comparable. Perhaps in the warm-season, the forecast precipitation amount is very sensitive to the vagaries of the convective parameterization and its triggering scheme; if the parameterization is not uniformly accurate, then the extra predictor, precipitable water, can provide useful information.

When 2-m temperatures and 10-m winds were included as predictors into the basic analog technique, the skill was uniformly poorer than the basic analog technique (Fig. 12). While it is possible that temperatures and winds may have some predictive value in some circumstances, in this case the pre-specified equal weightings of precipitation, temperature, and wind components was not a good choice; the de-emphasis of forecast precipitation as a predictor lessened the skill. However, a potential advantage of choosing analogs by a multivariate fit of precipitation, temperature, and winds is that if some users desired information on joint probabilities (how likely is it to be cold, windy, and wet?), the joint probability distribution could have been estimated directly from this set of analogs, while analogs chosen by a closeness of fit of precipitation forecasts would likely not be of much use for estimating winds, temperatures, or joint distributions.

As indicated in section 3.b.8, one deficiency of the basic analog technique that would be desirable to correct was a tendency for under-forecasting precipitation probabilities, especially at short leads (Fig. 13a). This was due to the skewed, often exponentially shaped climatological pdf of forecast precipitation, causing a bias in the selection of closest analogs toward those with less forecast amounts. When the rank analog technique was used, the reliability was markedly improved (Fig. 13b), the BSS was also substantially higher for the 2.5 mm forecasts, especially at the short forecast

leads, and the skill improvement was consistent across seasons (Fig. 14). However, the 25-mm rank-analog forecasts were slightly less skillful than the basic analog. The rank analog forecasts were more reliable (not shown), but they were slightly less sharp.

Would forecasts be improved if the rank analog technique used a smaller search region than the 16 points in Fig. 3? When using the inner 2x2 grid points, at short leads the skill of the 25-mm forecasts was improved substantially relative to the rank analog technique (Fig. 15). However, for the lighter precipitation amounts, the longer-lead forecasts in the warm season were slightly worse when using the smaller search region. The improvement at short leads for the high precipitation threshold indicated that the important predictor was the local precipitation forecast, and matching the pattern in the larger surrounding region was a less important consideration. The reason for the shorter-term forecasts being improved with the smaller search region was that systematic errors of the position bias were much smaller for the shorter-range forecasts. A cross-correlation analysis was performed that determined which NARR grid point had the most highly rank-correlated precipitation analysis with a given forecast grid point's ensemble-mean precipitation during the summer. Averaged over the conterminous U.S., 4.85 grid points separated the highest-correlated observed location from the original grid point for a 1-day forecast, and 8.64 grid points for a 5-day forecast.

Finally, consider the effect of the smoothing algorithm discussed in section 3.b.10. This technique was the same as the rank analog technique, but now the probability forecasts were smoothed to eliminate discontinuities in the probabilities along box boundaries. The smoothing produced a very slight improvement in the 2.5-mm forecast skill but improved the 25-mm forecast skill more substantially, especially at

short leads (Fig. 16). An example of the subtle effects of the smoothing are shown in Fig. 17. Notice that the probability discontinuities in central Tennessee and western Georgia were smoothed between panels a and b.

5. **Conclusions and discussion.**

In this article we have examined the skill of 24-h accumulated probabilistic quantitative precipitation forecasts from a variety of analog techniques that utilized a new, 25-year global reforecast data set produced by NOAA.

A general theory for probabilistic weather forecasting based on analogs was first proposed. Suppose an estimate is sought for the observed state's pdf given today's numerical forecast. Suppose also that we were provided with a nearly infinite set of reforecasts (hindcasts) and associated observations and that the climate was stable. Then, past model forecast states could be identified that are nearly identical to the current forecast state. Given the dates of these past analog forecasts, the asymptotically correct probabilistic forecast can be formed from the distribution of observed states on those dates.

This general theory could not be applied to global weather prediction given a limited set of reforecasts, for the chance of finding even one similar forecast analog in that limited set is highly improbable. However, approximations can be made to this theory to make it useful for statistically correcting weather forecasts. For instance, when estimating the local pdf of the observed state given the forecast, it was possible to choose the forecast analogs only based on the local weather. There commonly is an ample

supply of highly similar local forecasts given a modest-length reforecast to compare. However, the rarer the event, the more difficult it is to find close forecast analogs.

We then examined several approximate PQPF analog forecast techniques, using a 25-year set of ensemble reforecasts. The analog techniques found past ensemble-mean forecasts in a local region that were similar to today's ensemble-mean forecasts in that region and formed probabilistic forecasts from the observed weather on the dates of the past analogs. All of the analogs techniques provided dramatic improvements in the Brier skill score relative to basing probabilities on the raw ensemble counts or the counts corrected for bias. However, the analog techniques were generally similar in skill to those from a logistic regression technique. Comparing the various analog techniques tested, we found the following: (1) Finding analogs for each member rather than for the ensemble mean generally decreased the forecast skill. (2) Finding analogs by matching not only mean forecast precipitation but also mean forecast precipitable water improved short-range, warm-season forecasts. (3) Finding analogs by matching surface winds and temperatures in addition to precipitation decreased the precipitation forecast skill. (4) Finding analogs based on the closeness of the relative rank of the mean forecast rather than its magnitude improved reliability at the short forecast leads. (5) A smaller search region was preferable when finding analogs for short-range forecasts, and a larger search region was preferable for longer-range forecasts. (6) Smoothing increased the skill of the forecasts slightly.

We also considered the effectiveness of a proposed bias correction technique that adjusted precipitation amounts so that over many cases the forecast cumulative density function would match the observed cumulative density function. This procedure has

been used in a slightly modified form at NCEP since 2004. This technique tended to improve forecast skill relative to the raw ensemble in the wintertime but worsen it in the summer. We determined that a CDF correction was generally unwise when the forecast and observed data are not highly correlated.

Despite the demonstrated skill and reliability of the reforecast-based techniques, many may believe it unwise to utilize forecast products from a T62, 1998 version of the NCEP GFS. Wouldn't it be wiser to base a PQPF forecast upon raw output from more recent, higher-resolution model forecasts? While numerical precipitation forecast skill undoubtedly has improved in the past ten years, there is reason to believe that the analog reforecast product demonstrated here are still competitive with these much newer, higher-resolution forecast models (HWM05). Calibrated products based on future, higher-resolution reforecasts should be even more competitive, for even with a better model, calibration with reforecasts still has been shown to provide substantial benefit (Whitaker and Vitart 2005).

For reforecasts to be of most benefit, the current numerical forecast should be conducted with the same model and data assimilation methods used in the production of the reforecast. This of course requires freezing the forecast model. Considered in isolation, this approach would be unattractive to weather prediction facilities, which would prefer to implement forecast model improvements quickly. Perhaps a dual-track system can be used, whereby an inexpensive fixed, reduced-resolution version of the forecast model is run alongside the frequently upgraded, operational higher-resolution version. Users can choose for themselves whether they'd prefer guidance from the statistically adjusted older model, raw guidance from the newer model, or some blend.

Perhaps every few years, a new reforecast data set would be produced with a more recent, higher-resolution version of the model, so that the calibrated probabilistic guidance could leverage the improvement in the forecast models.

The literature has yet to demonstrate that quantum jumps in skill can be achieved with small training data sets. The skill increases in precipitation forecasts we have demonstrated here are equivalent to the skill increases afforded by many years of sustained model development by a large staff of scientists. While we have concentrated here on demonstrating a calibration technique for precipitation, the statistical problems with precipitation are likely to be much more difficult than, say, for other commonly desired weather elements such as surface temperature. We expect that with a long reforecast data set (saving more forecast variables than we did for this pilot project), it should be possible to produce calibrated probabilistic forecasts for even the thorniest of problems, such as precipitation type or severe-weather probability.

The US National Weather Service is currently considering how to make skillful, reliable probabilistic weather forecasts a part of its National Digital Forecast Database (Glahn and Ruth 2003, Mass 2003ab, Glahn 2003, 2005, Abrams 2004). Perhaps reforecast-based techniques are the most straightforward and promising way to achieve this goal.

Acknowledgments

The first author's participation was partially supported by NSF grant ATM-0130154.

References

Abrams, E., 2004: Implementation and refinement of digital forecasting databases. *Bull. Amer. Meteor. Soc.*, **85**, 1667-1672.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, **12**, 581-594.

Gallus, W. A., Jr., 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting*, **17**, 1296-1302.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.

----------, and D. P. Ruth, 2003: The digital forecast database of the National Weather Service. *Bull. Amer. Meteor. Soc.*, **84**, 195-201.

---------- , 2003: Comments on "IFPS and the future of the National Weather Service." *Wea. Forecasting*, **18**, 1299-1304.

---------- , 2005: Comments on "Implementation and refinement of digital forecasting databases." *Bull. Amer. Meteor. Soc.*, **86**, 1315-1318.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

----------, J. S. Whitaker, and X. Wei, 2004: Ensemble re-forecasting: improving medium range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434-1447.

---------- , ---------- , and S. L. Mullen, 2005: Reforecasts, an important new data set for improving weather predictions. *Bull. Amer. Meteor. Soc*., in press. Available at http://www.cdc.noaa.gov/people/tom.hamill/reforecast_bams4.pdf .

---------- , and J. Juras, 2005: Overestimating forecast skill through improper applications of verification metrics: Simpson's paradox in meteorology. *Mon. Wea. Rev*., in review. Available at www.cdc.noaa.gov/people/tom.hamill/skill_overforecast_v2.pdf.

Hastie, T. J., and R. J. Tibshirani, 1990: *Generalized Additive Models*. Chapman and Hall, 335 pp.

Kalnay, E., and co-authors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc*., **77**, 437-472.

Lorenz, E. N., 1993: *The Essence of Chaos*. University of Washington Press, 227 pp.

Mass, C. F., 2003: IFPS and the future of the National Weather Service. *Wea. Forecasting*, **18**, 75-79.

---------- , 2003: Reply. *Wea. Forecasting*, **18**, 1305-1306.

Mesinger, F., and coauthors, 2005: North American regional reanalysis. *Bull. Amer. Meteor. Soc*., submitted. Available from fedor.mesinger@noaa.gov.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran* (2nd Ed.). Cambridge Press, 963 pp.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc*. **74**, 2317-2330.

----------, and ---------- , 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev*., **125**, 3297-3319.

Van den Dool, H. M., 1994: Searching for analogues, how long must we wait?, *Tellus*, **45A**, 314-324.

----------, 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev*., **117**, 2230-2247.

Whitaker, J. S., X. Wei., and F. Vitart, 2005: Improving week-two forecasts with multi-model re-forecast ensembles. *Mon. Wea. Rev*., accepted. Available from Jeffrey.s.whitaker@noaa.gov.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Cambridge Press, 467 pp.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods, *J. Climate*, **12**, 2474-2489.

List of Figures

Figure 1: Illustration of two methods for setting probabilities using synthetic reforecast data. Dots are the reforecast data (abscissa) and the associated observed value (ordinate). Vertical bars denote bins for considering "nearby" analogs. Heavy horizontal solid lines are the probabilities set by relative frequency in these bins (axis labels on right). Dashed line is the probability set by logistic regression.

Figure 2: Illustration of the bias-correction technique described in section 3.b.2. Dashed line denotes the observed CDF, solid line the forecast CDF. A raw forecast of 7 mm is at approximately the 91st percentile of the forecast CDF; the 91st percentile of the observed CDF is approximately 5.6 mm. Thus, the precipitation forecast is changed from 7 to 5.6 mm.

Figure 3: Illustration of the basic analog technique for a 2-day forecast. The ensemble mean precipitation forecast is shown in the first row, defined at the 16 dots. Analogs and probability forecasts are desired for the dashed box in the middle. The four closest matching 2-day ensemble-mean forecasts are shown in the second row, and the higher-resolution observed weather on those dates are shown in the third row. Probabilistic forecasts formed from the observed analogs are shown in the fourth row for 3, 10, and 25 mm thresholds, and the observed data is shown in the bottom row.

Figure 4: Illustration of the smoothing algorithm. Probabilities are sought in this case for the NARR grid points colored orange. The nine panels are the nine regions where

analog dates have been calculated that overlap the orange grid points. Other NARR grid points are denoted by small black dots. Analog matches are calculated by forecast similarity at the large black dots; the center of each analog search region is denoted by the blue dot. For the orange grid point highlighted by the red box, the final probability is a weighted sum of the probabilities of the nine estimates, the weight being determined by the relative distance of the blue dot in that panel from the red box.

Figure 5: Brier skill score of the ensemble relative frequency technique as a function of the time of the year and the lead time of the forecast. (a) Skill at 2.5 mm, (b) skill at 25 mm.

Figure 6: As in Fig. 5, but for the bias-corrected relative frequency technique.

Figure 7. (a) Illustration of forecast and observed CDFs for 1-day forecasts in northern Mississippi during August. (b) Scatter plot of one ensemble member's forecast vs. observed forecast. Red curve illustrates the remapping that will occur between a forecast precipitation amount and the corrected amount, based on the CDF correction technique. Orange curve denotes remapping between forecast and mean observed given the forecast. (c) For 10 mm, a typical ensemble forecast distribution with a mean of 23-27 mm (solid line), the adjusted distribution (dashed line), and the conditional distribution of the observed values given an ensemble mean of 23-27 mm (in red).

Figure 8: BSS of the basic analog technique as a function of the month and the lead time of the forecast. (a) Skill at 2.5 mm, (b) skill at 25 mm.

Figure 9: Monthly BSS of the logistic regression technique, as in Fig. 8. Grey shading on forecasts indicates skill difference relative to basic analog approach in Fig. 8; grey shading below the reference line indicates more skill than the basic analog approach, and shading above the reference line indicates less skill.

Figure 10: Monthly BSS of the basic technique using individual members. Skill differences (grey shading, as in Fig. 9) are relative to the basic analog technique in Fig. 8.

Figure 11: Monthly BSS of the basic technique including precipitable water, with skill again compared via shading relative to the basic analog technique in Fig. 8.

Figure 12: Monthly BSS of the basic technique including 2-m temperatures and 10-m wind but for the basic technique including 2-m temperature and 10-m winds, with skill again compared via shading relative to the basic analog technique in Fig. 8.

Figure 13: Reliability diagrams for 2.5 mm 1-day forecasts from (a) 50-member basic analog technique, and (b) 50-member rank analog technique.

Figure 14: Monthly BSS for the rank analog technique, with skill again compared via shading relative to the basic analog technique in Fig. 8.

Figure 15: Monthly BSS of the rank analog with smaller search region technique. Skill differences here are with respect to the rank analog technique in Fig. 14.

Figure 16: Monthly BSS of the smoothed rank analog technique. Skill differences here are with respect to the rank analog technique in Fig. 14.

Figure 17: Probability of greater than 2.5 mm precipitation for the 24-h period starting 0000 UTC 11 January 1994, from (a) rank analog technique, and (b) smoothed rank analog technique.

Technique	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
1) Ensemble relative frequency	.0840	-.0486	-.1098	-.1624	-.2117	-.2552
2) Bias-corrected relative frequency	.2642	.1753	.0597	-.0424	-.1318	-.2033
3) Basic analog	.4026	.3443	.2648	.1923	.1335	.0853
4) Logistic regression	.4108	.3395	.2564	.1842	.1266	.0815
5) Basic using individual members	.4061	.3414	.2555	.1774	.1155	.0692
6) Basic including precipitable water	.4080	.3486	.2687	.1969	.1378	.0898
7) Basic including 2-m temperature and 10-m winds	.3803	.3312	.2565	.1881	.1319	.0875
8) Rank analog	.4195	.3555	.2726	.1965	.1360	.0865
9) Rank analog with smaller search region	.4194	.3496	.2635	.1871	.1272	.0791
10) Smoothed rank analog	.4260	.3613	.2779	.2020	.1415	.0925
Difference that's statistically significant, 2-sided test, $\alpha = 0.05$.	.0010	.0009	.0008	.0007	.0006	.0006

Table 1: Brier Skill Score for various forecast techniques at 2.5 mm, averaged over the 25 years. Last row provides the amount of difference between two forecasts that is considered statistically significant according to a 2-sided test with a=0.05. Highest score for a particular day in boldface type.

Technique	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
1) Ensemble relative frequency	.0534	-.0668	-.0624	-.0473	-.0535	-.0559
2) Bias-corrected relative frequency	.0105	-.0503	-.0684	-.0731	-.0860	-.0894
3) Basic analog	.1816	.1298	.0887	.0597	.0357	.0201
4) Logistic regression	.1895	.1205	.0831	.0572	.0350	.0219
5) Basic using individual members	.1856	.1267	.0815	.0504	.0278	.0131
6) Basic including precipitable water	.1841	.1319	.0903	.0607	.0370	.0212
7) Basic including 2-m temperature and 10-m winds	.1715	.1245	.0854	.0587	.0363	.0217
8) Rank analog	.1727	.1260	.0878	.0588	.0350	.0193
9) Rank analog with smaller search region	.1860	.1280	.0865	.0568	.0326	.0176
10) Smoothed rank analog	.1832	.1318	.0912	.0621	.0378	.0222
Difference that's statistically significant, 2-sided test, $\alpha = 0.05$.	.0015	.0012	.0010	.0009	.0007	.0007

Table 2: As in Table 1, but for 25 mm.

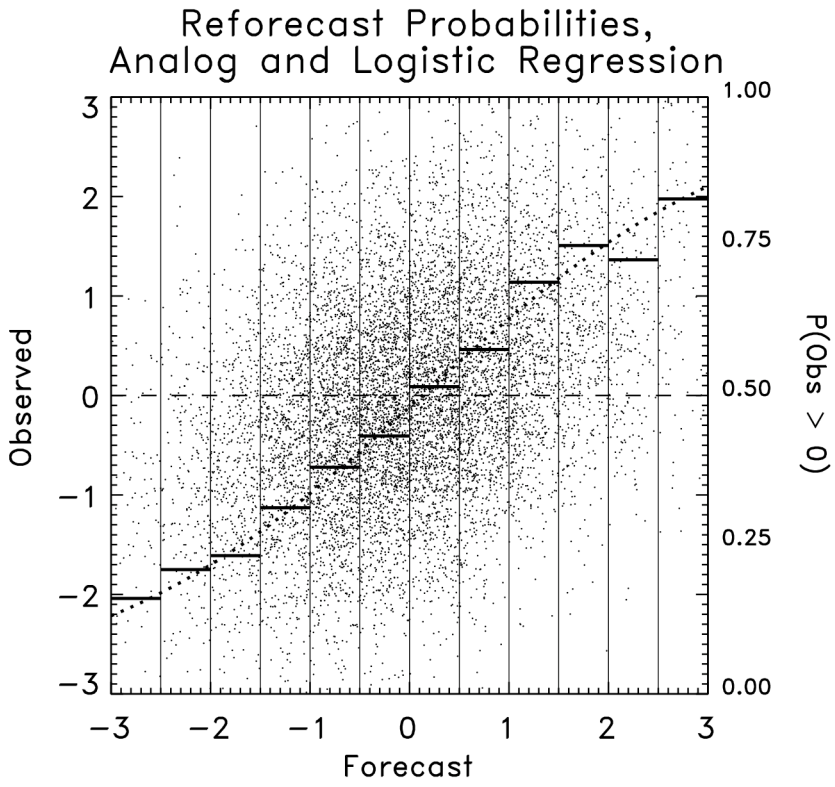

Figure 1: Illustration of two methods for setting probabilities using synthetic reforecast data. Dots are the reforecast data (abscissa) and the associated observed value (ordinate). Vertical bars denote bins for considering "nearby" analogs. Heavy horizontal solid lines are the probabilities set by relative frequency in these bins (axis labels on right). Dashed line is the probability set by logistic regression.

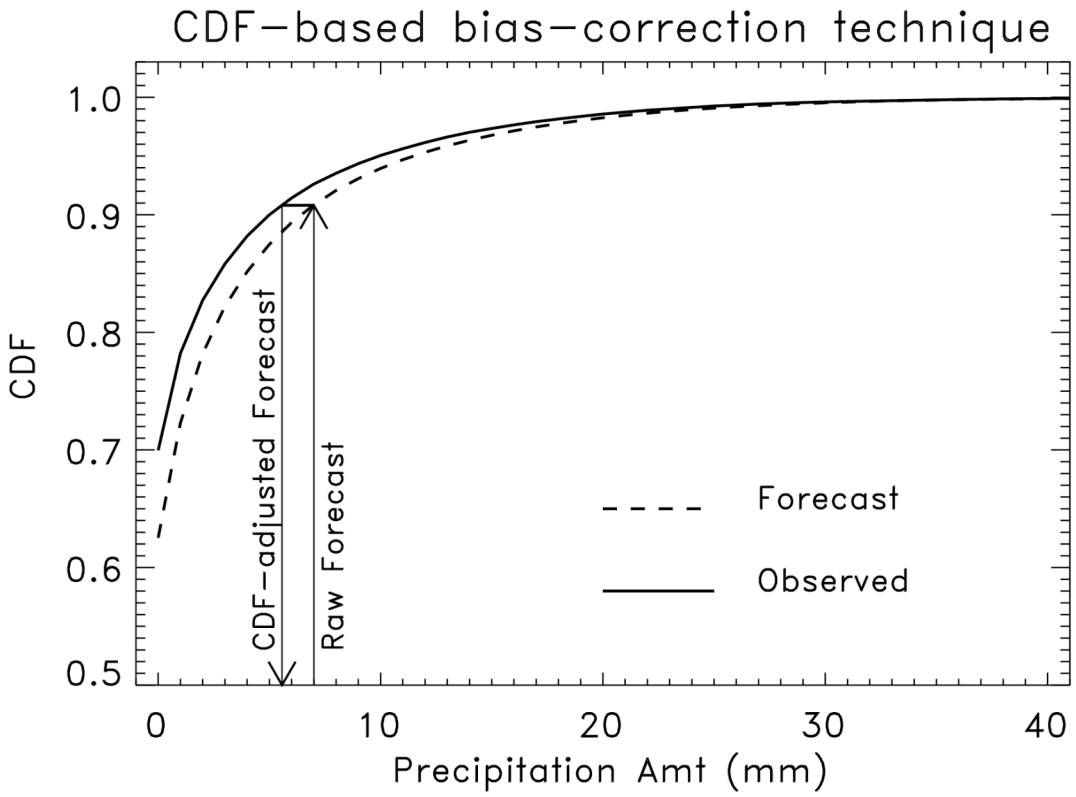

Figure 2: Illustration of the bias-correction technique described in section 3.b.2. Dashed line denotes the observed CDF, solid line the forecast CDF. A raw forecast of 7 mm is at approximately the 91st percentile of the forecast CDF; the 91st percentile of the observed CDF is approximately 5.6 mm. Thus, the precipitation forecast is changed from 7 to 5.6 mm.

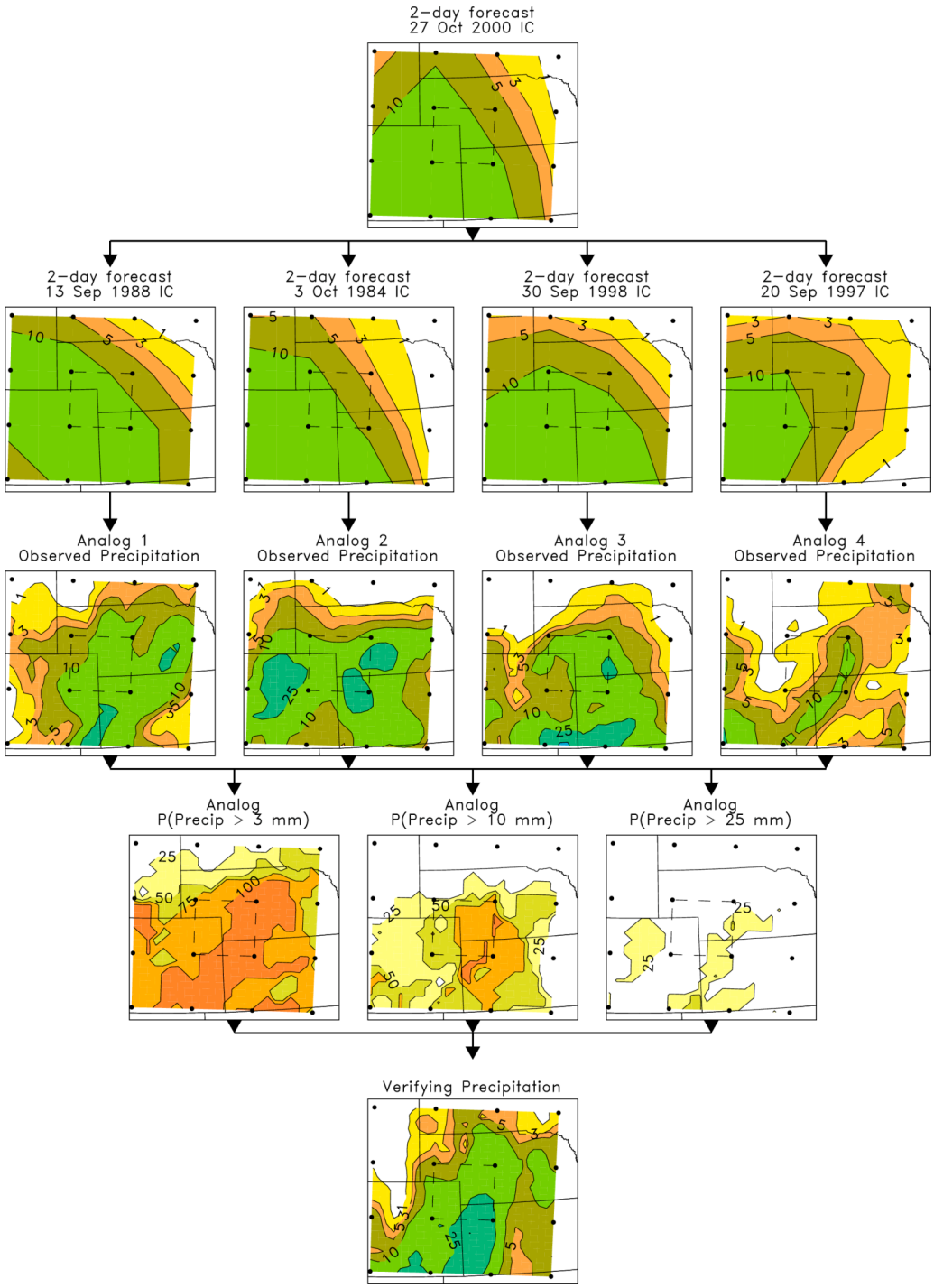

Figure 3: Illustration of the basic analog technique for a 2-day forecast. The ensemble mean precipitation forecast is shown in the first row, defined at the 16 dots. Analogs and probability forecasts are desired for the dashed box in the middle. The four closest matching 2-day ensemble-mean forecasts are shown in the second row, and the higher-resolution observed weather on those dates are shown in the third row. Probabilistic forecasts formed from the observed analogs are shown in the fourth row for 3, 10, and 25 mm thresholds, and the observed data is shown in the bottom row.

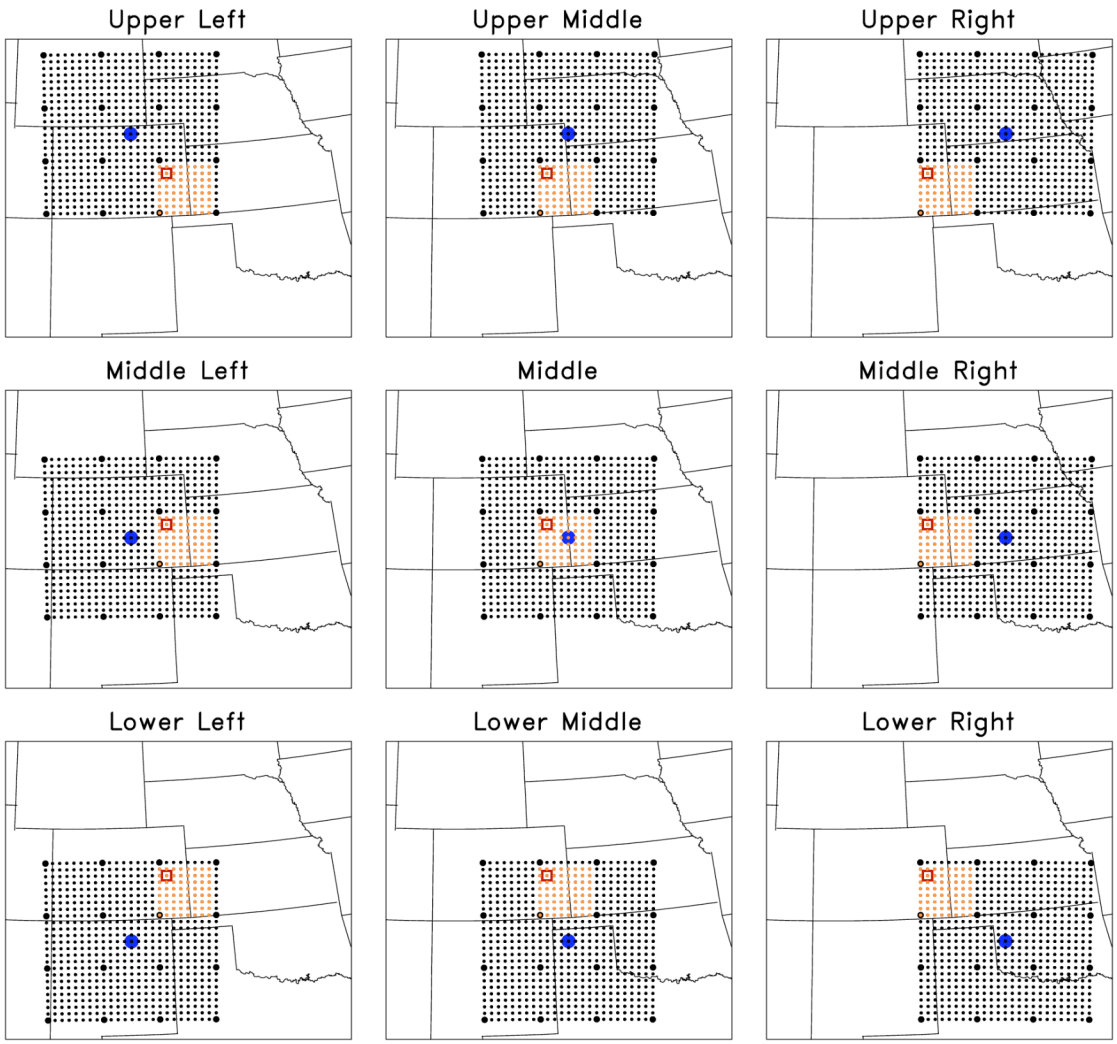

Figure 4: Illustration of the smoothing algorithm. Probabilities are sought in this case for the NARR grid points colored orange. The nine panels are the nine regions where analog dates have been calculated that overlap the orange grid points. Other NARR grid points are denoted by small black dots. Analog matches are calculated by forecast similarity at the large black dots; the center of each analog search region is denoted by the blue dot. For the orange grid point highlighted by the red box, the final probability is a weighted sum of the probabilities of the nine estimates, the weight being determined by the relative distance of the blue dot in that panel from the red box.

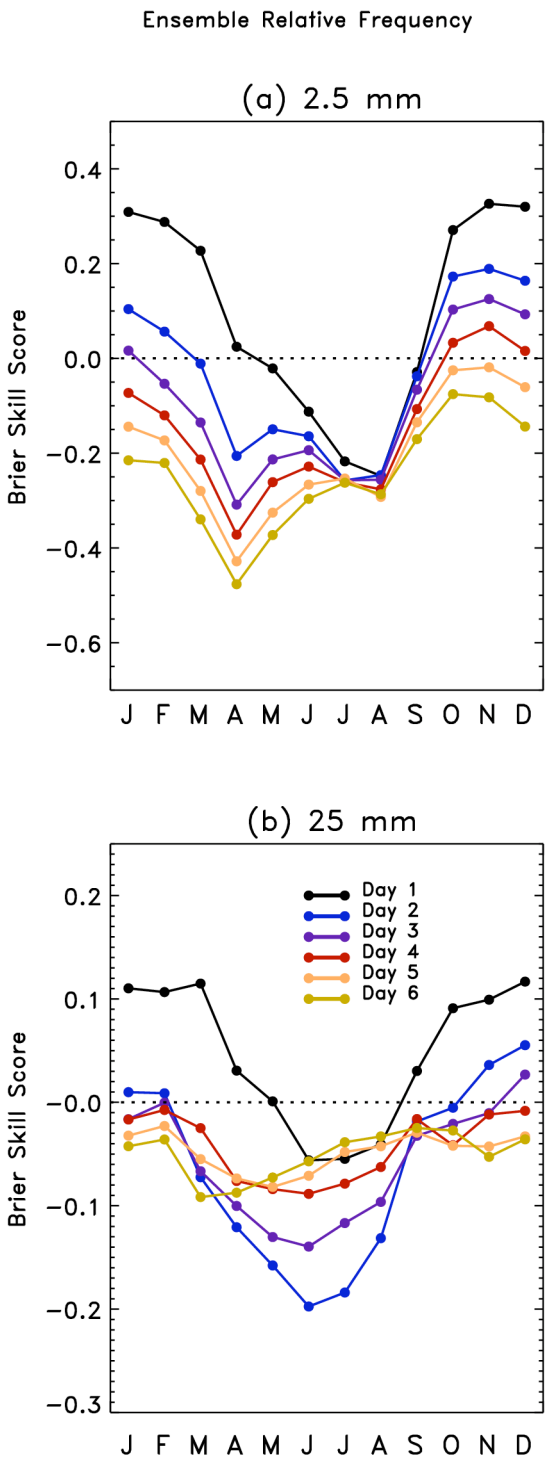

Figure 5: Brier skill score of the ensemble relative frequency technique as a function of the time of the year and the lead time of the forecast. (a) Skill at 2.5 mm, (b) skill at 25 mm.

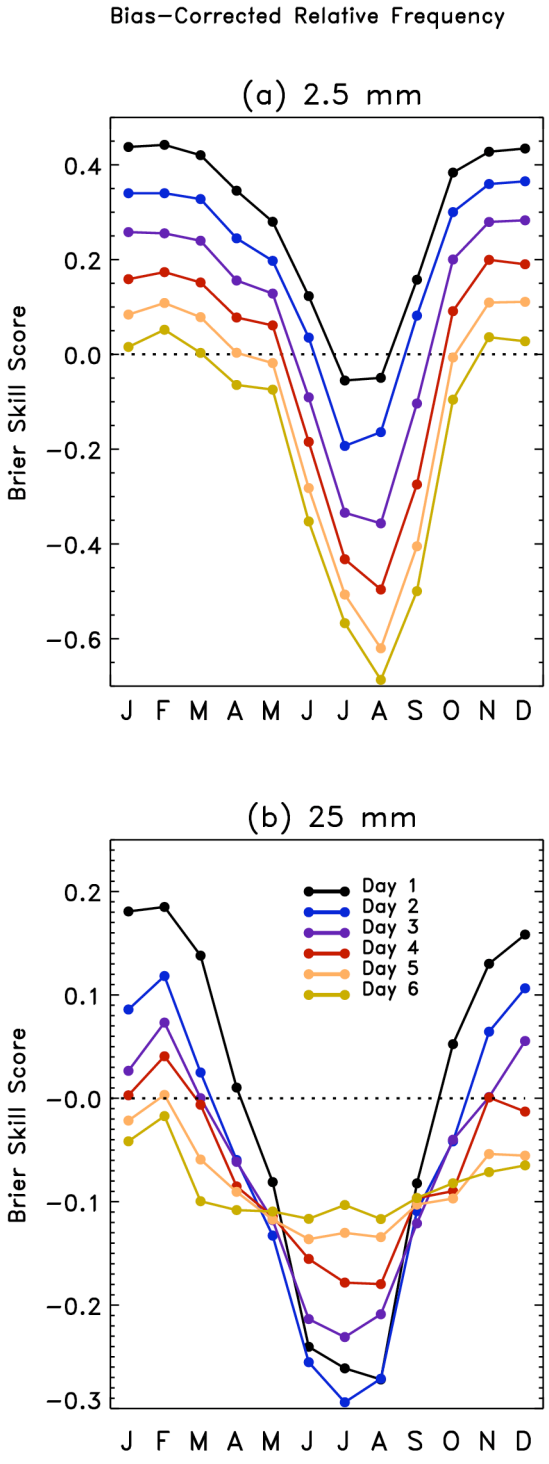

Figure 6: As in Fig. 5, but for the bias-corrected relative frequency technique.

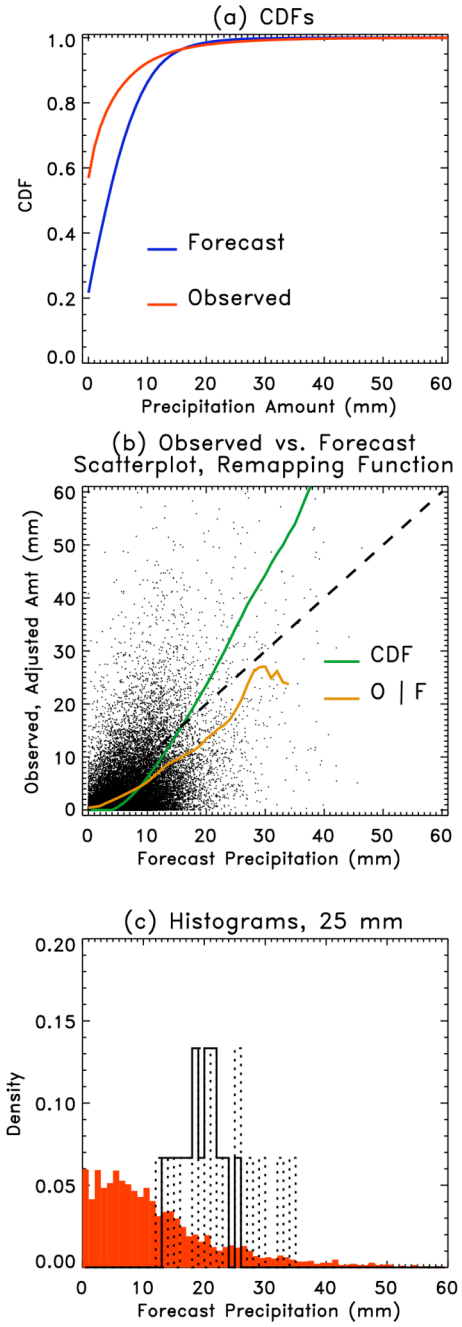

Figure 7. (a) Illustration of forecast and observed CDFs for 1-day forecasts in northern Mississippi during August. (b) Scatter plot of one ensemble member's forecast vs. observed forecast. Red curve illustrates the remapping that will occur between a forecast precipitation amount and the corrected amount, based on the CDF correction technique. Orange curve denotes remapping between forecast and mean observed given the forecast. (c) For 10 mm, a typical ensemble forecast distribution with a mean of 23-27 mm (solid line), the adjusted distribution (dashed line), and the conditional distribution of the observed values given an ensemble mean of 23-27 mm (in red).

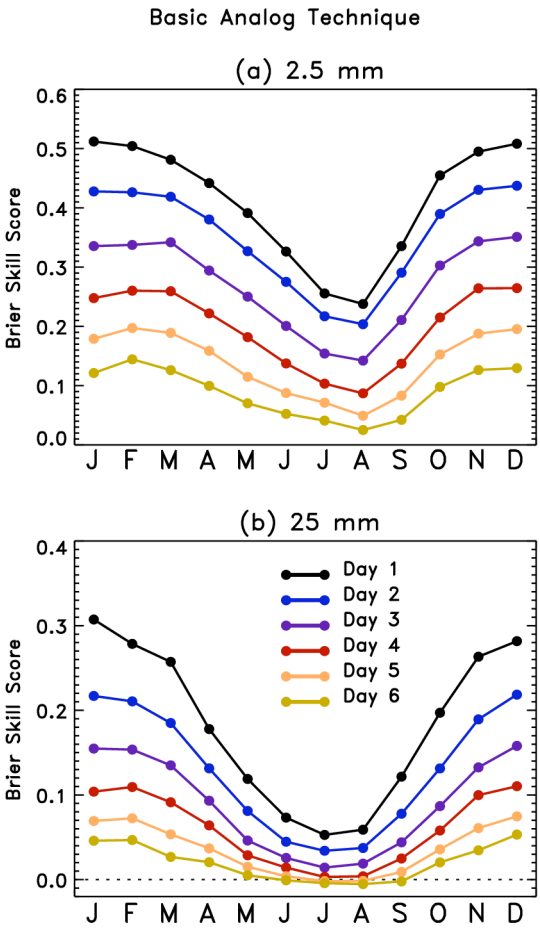

Figure 8: BSS of the basic analog technique as a function of the month and the lead time of the forecast. (a) Skill at 2.5 mm, (b) skill at 25 mm.

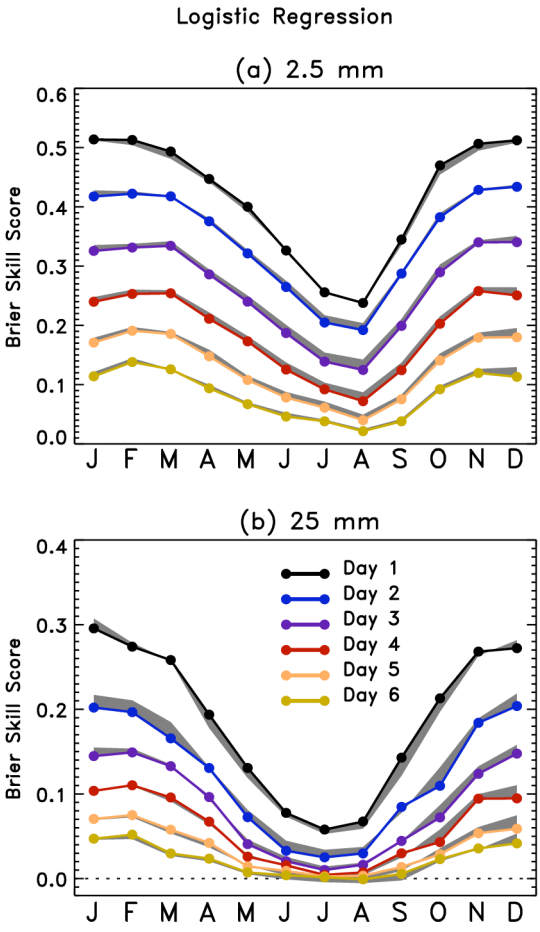

Figure 9: Monthly BSS of the logistic regression technique, as in Fig. 8. Grey shading on forecasts indicates skill difference relative to basic analog approach in Fig. 8; grey shading below the reference line indicates more skill than the basic analog approach, and shading above the reference line indicates less skill.

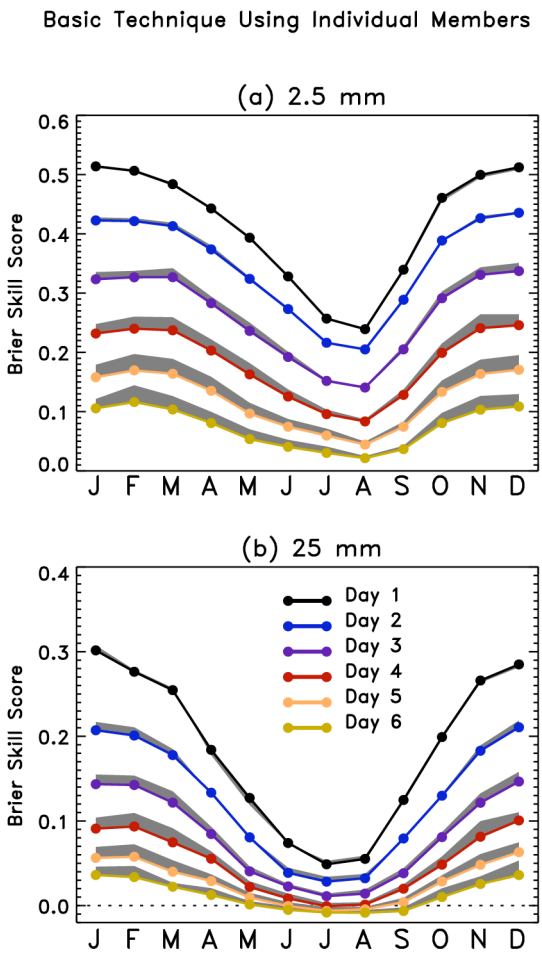

Figure 10: Monthly BSS of the basic technique using individual members. Skill differences (grey shading, as in Fig. 9) are relative to the basic analog technique in Fig. 8.

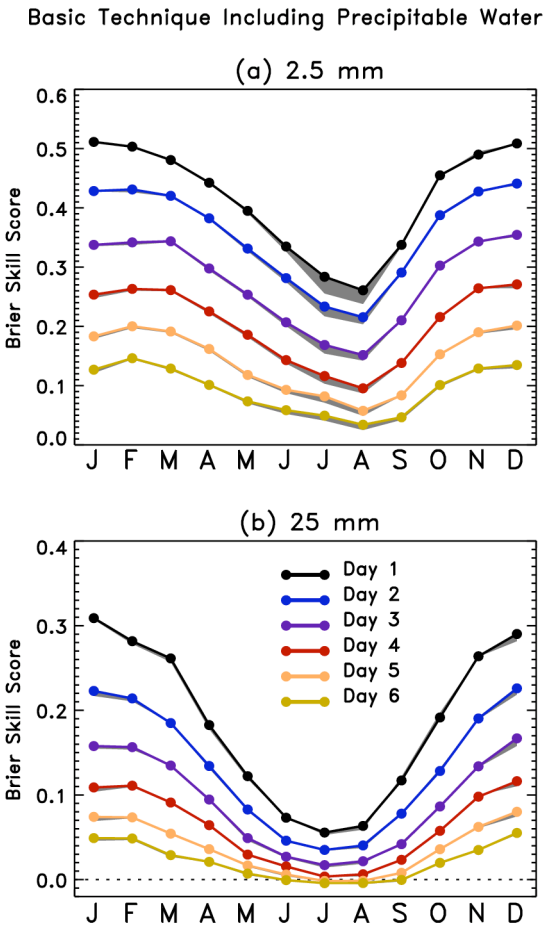

Figure 11: Monthly BSS of the basic technique including precipitable water, with skill again compared via shading relative to the basic analog technique in Fig. 8.

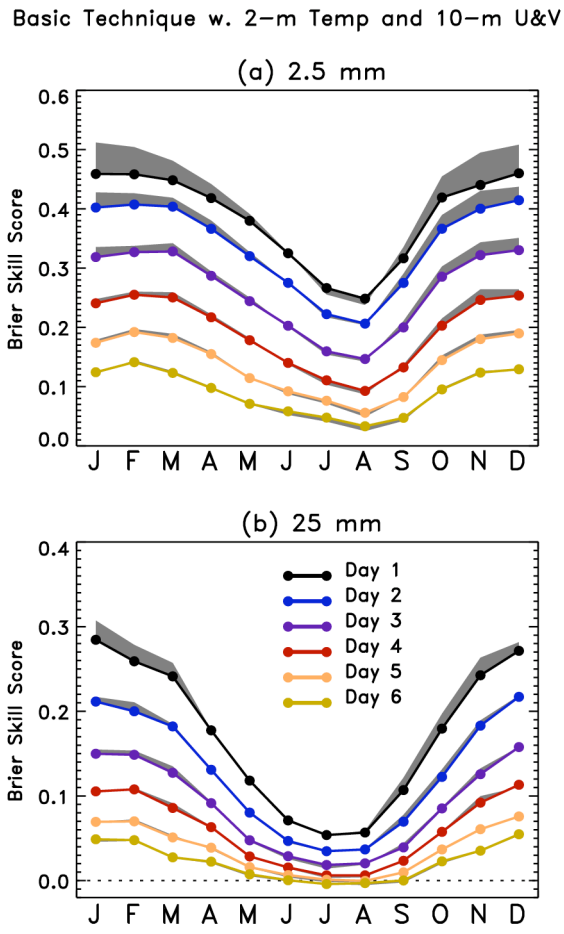

Figure 12: Monthly BSS of the basic technique including 2-m temperatures and 10-m wind but for the basic technique including 2-m temperature and 10-m winds, with skill again compared via shading relative to the basic analog technique in Fig. 8.

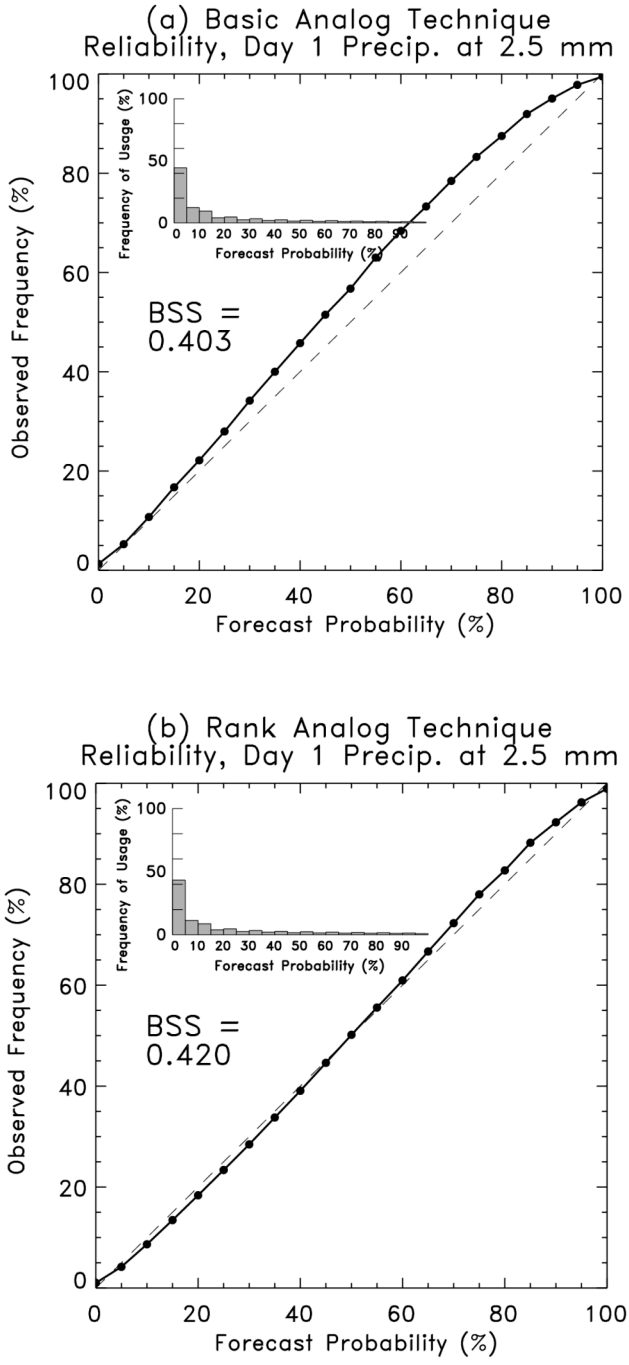

Figure 13: Reliability diagrams for 2.5 mm 1-day forecasts from (a) 50-member basic analog technique, and (b) 50-member rank analog technique.

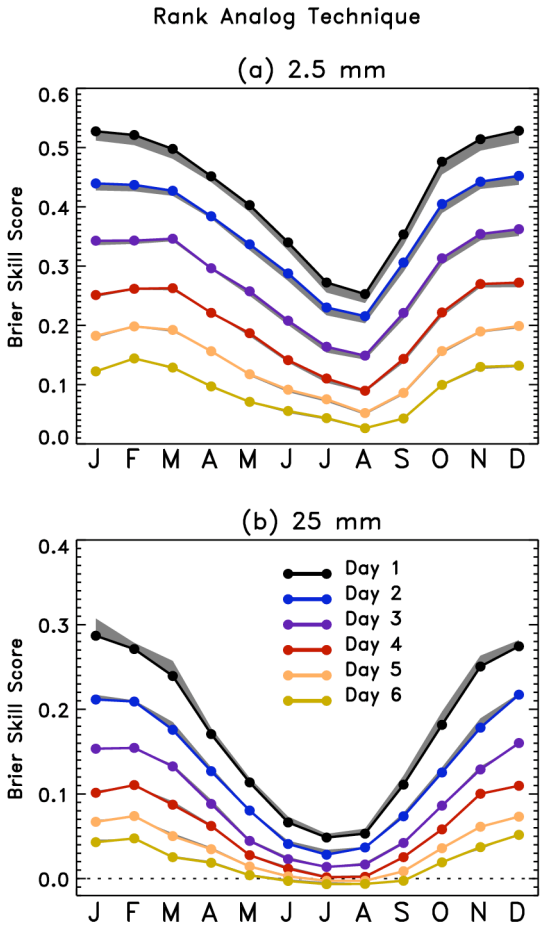

Figure 14: Monthly BSS for the rank analog technique, with skill again compared via shading relative to the basic analog technique in Fig. 8.

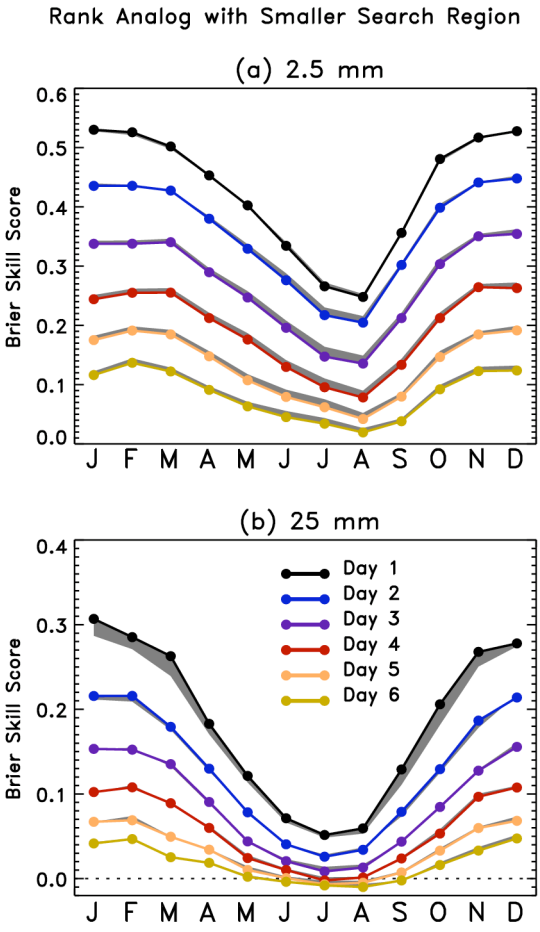

Figure 15: Monthly BSS of the rank analog with smaller search region technique. Skill differences here are with respect to the rank analog technique in Fig. 14.

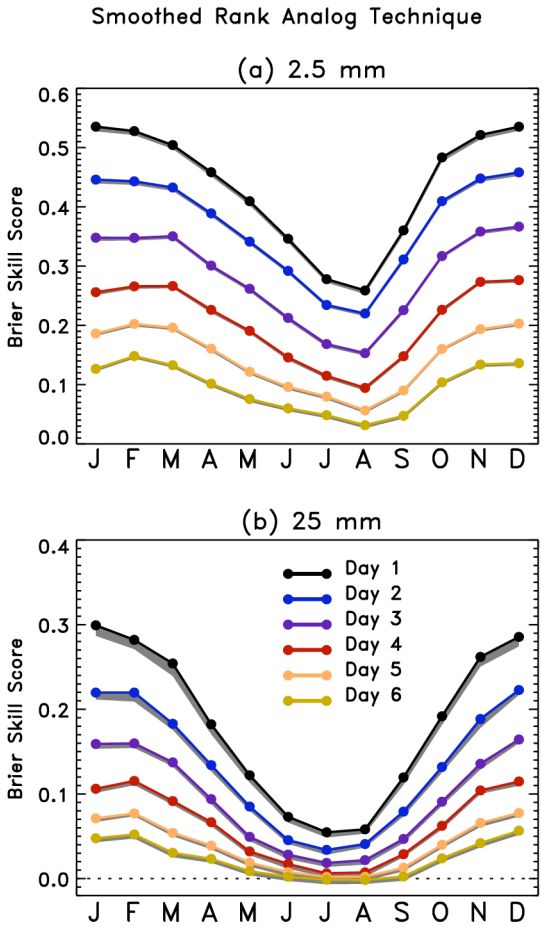

Figure 16: Monthly BSS of the smoothed rank analog technique. Skill differences here are with respect to the rank analog technique in Fig. 14.

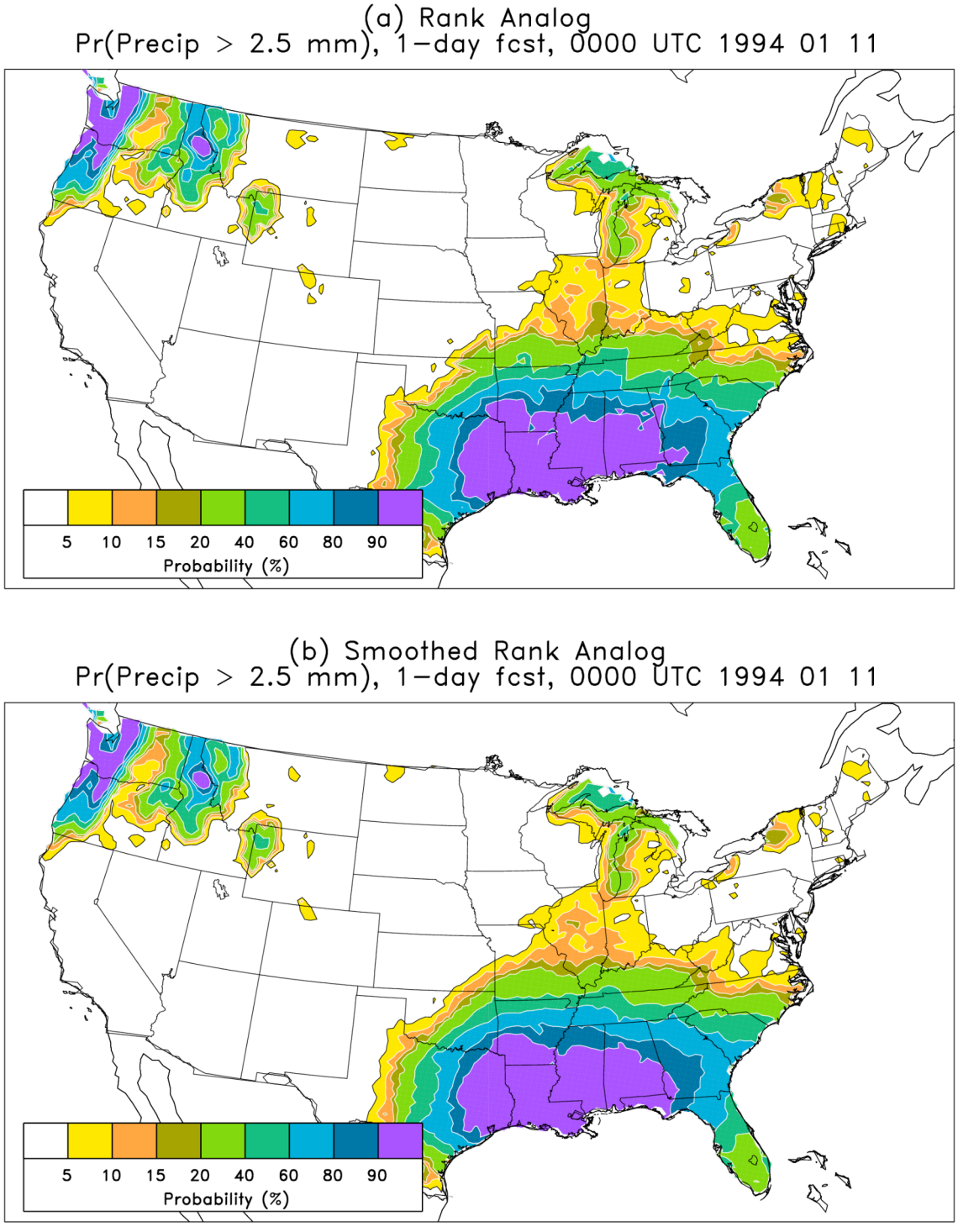

Figure 17: Probability of greater than 2.5 mm precipitation for the 24-h period starting 0000 UTC 11 January 1994, from (a) rank analog technique, and (b) smoothed rank analog technique.